Data Preparation and Quality Challenges for the Personality Recognition in Indian Languages using Machine Learning and Deep Learning Approaches

Jayshri Patil[1], **Jikitsha Sheth**[2]

[1]BMCCA, Bhagwan Mahavir University, Surat, Gujarat, India
[2]SRIMCA, Uka Tarsadia University, Bardoli, Gujarat, India

E-mail: [1]jayshri.patil@bmusurat.ac.in, [2]jikitsha.sheth@utu.ac.in

Abstract

Information about the user and their feelings, thoughts, and emotions are expressed through the status, comments, and updates on social media or other platforms. These user-generated contents are an important source for recognizing a user's personality. Due to the increase in the amount of various Indian language contents on social media, there is a necessity to recognize personality from Indian languages. The challenges have increased in the collection and generation of datasets due to the lack of resources for Indian languages. In the field of personality recognition, the researchers have utilized machine learning and deep learning techniques to infer users' personalities. The machine learning and deep learning models require enough labeled data for the training. Unlike traditional machine learning, deep learning techniques automatically generate features and require a significant amount of labeled data. For the personality recognition task from the Indian language, no sufficient annotated dataset is available and data preparation for the personality recognition task in the language has become a critical issue. This paper represents the existing gold standard dataset for personality recognition in English and also focuses on the challenges of a large amount of labeled data preparation in the Indian language.

Keywords: Personality, Personality Recognition, Machine Learning, Deep Learning

1. Introduction

Individuals express themselves through utterances and writing with their unique style. The researcher can recognize individual identity or personality to some extent in the way

DOI: https://doi.org/10.36548/jismac.2022.1.004
Received: 22.03.2022, received in revised form: 13.04.2022, accepted: 01.05.2022, published: 13.05.2022

people talk and write [12]. The literature [3,4,11,15] has documented expressive writing techniques where authors engage in deep and meaningful writing about a troubling, stressful, and emotional event. The person writing samples or speech samples plays a key role in the recognition of their personality.

The researchers in the field of personality recognition have used machine learning and deep learning approaches. From the literature survey, it has been concluded that deep learning is the fast-growing approach in the field of personality recognition and it improves accuracy rates in the results [1,2,8,9,11,14,16,18]. Deep learning models automatically generate features and require large amounts of training data to perform well. The gold-standard datasets are available in English for the personality recognition task and state-of-the-art computational personality recognition methods are performed on these datasets. The following section describes the existing datasets available in English for the personality recognition task.

2. Existing Dataset for the Personality Recognition Task

The users' utterances words convey a great deal of information about them. For the personality recognition task, the researchers have collected datasets either from social media platforms or even from individuals who are asked to write an essay.

2.1 Written text or Conversation Extracts

The Stream-of-consciousness Essays dataset is a large dataset written by psychology students who were said to write whatever thoughts and feelings comes into their mind for 20 minutes. It comprises 2,479 essays with 1.9 million words and these essays are tagged with personality traits based on the Big-5 personality test given by the students. This data was collected and analyzed by the authors in [12]. This dataset has been utilized in research work [1,4,11]. Another dataset collected by the authors in [21] consists of 96 participants' conversation extracts recorded using an Electronic Activated Recorder (EAR). It includes 97,468 words and 15,269 utterances.

2.2 Social Media

The use of social media is increasing tremendously in this internet era. Users share their information, views, and emotion through social media platforms. The researchers in the work [5-9] have utilized data collected from Facebook and Twitter to infer users' personalities. The user-generated content on social network platforms Facebook and Twitter was collected through the myPersonality project. Collecting labeled data from social media is a time-consuming and expensive process. This is the reason for not many standard datasets from social media are available for the personality recognition task [7]. For collecting personality scores two approaches have been utilized by the researchers. First, participation of users to provide self-reported personalities through answering questionnaires. The second approach is by asking other users' opinions on the personality of a user [7]. Labeling or assigning the personality score is a challenging task for a non-expert. In personality psychology, various questionnaires and tests are available to find personality scores such as Ten-Item Personality Inventory (TIPI), 10-item personality test (BFI-10), NEO-Personality-Inventory (240 items), the NEO Five-Factor Inventory (60 items), and the Big-Five Inventory (44 items) [12]. The utilization of these questionnaires makes it easier to collect the users` personality scores. The researcher in [8,16] employed the Big Five Factor personality Inventory questionnaire for collecting personality scores.

Many peoples live in rural areas and hence they have the problem of understanding and speaking the English language. There is a also need to explore the Indian language content and get insight and evaluation of human perceptions expressed by the users. For the personality recognition task from the Indian language, there are no sufficient annotated datasets available. It is difficult to find or generate a new dataset and as well as cleaning and re-labeling of existing data. The following section describes the challenges faced while the preparation of the dataset for the personality recognition task in the Indian language.

3. Data Preparation and Quality Challenges for the Personality Recognition In Indian Languages

From the literature survey, it has been found that research efforts in personality recognition mostly deal with English text; no work has been reported for personality recognition in the Indian language. For personality recognition from the Indian languages, there is no sufficient annotated dataset available in Indian languages. The machine learning and deep learning models learn an internal representation of data to perform the task. Therefore these models require a significant amount of data to learn well. Data preparation is the major challenge for the personality recognition task. Data collection and preparation recently has become a critical issue in the field of personality recognition tasks in the Indian language.

In the field of personality recognition, finding a suitable dataset and data labeling assigns personality labels to the user-generated content itself becomes a challenge. Data labeling is important in all supervised learning approaches. Manual labeling of user personalities to user-generated content can be expensive. The collection of more data may not improve the model's accuracy due to its low quality. To achieve good accuracy results of the machine learning and deep learning models data need to be clean. The machine learning and deep learning models automatically learn and generate features, and infer the personality from their user-generated content, but it requires a large amount of training labeled data. For the development of the dataset, the data management issues are how to prepare a large personality dataset, how to perform traits class labeling, and how to improve the quality of existing personality datasets. To deal with such issues and challenges researchers need to understand the literature related to personality psychology, machine learning and deep learning approaches, and data management communities. The challenges in dataset preparation and data cleaning are presented in this paper based on the survey study [13,17]. The solutions to these challenges are discussed in the following sections.

3.1 Data Collection and Preparation

The data collection process for the personality recognition task comprises three major processes. First, data acquisition is the process of finding a suitable dataset in the Indian language for the training of machine learning and deep learning models. For the data collection and preparation in Indian languages, researchers either can manually collect the datasets or translate the existing dataset available in English such as Stream-of-consciousness Essays dataset and myPersonality for the personality recognition task as discussed in Section-2. These existing datasets are translated by the Google Translate interface [19]. After the translation, there is a need to manually clean and correct the translation with the re-arrangement of words.

The second process for data collection is data labeling, it is necessary for the supervised learning approaches. The manual labeling of personality data is expensive. In data labeling, one can utilize existing personality labels or can use various questionnaires available in personality psychology to find users' personality scores such as Ten-Item

Personality Inventory (TIPI), 10-item personality test (BFI-10), NEO-Personality-Inventory (240 items), the NEO Five-Factor Inventory (60 items), and the Big-Five Inventory (44 items) as discussed in Section 2.2. Finally, researcher can improve the quality of the existing dataset instead of collecting and preparing data from the beginning.

3.2 Improving Existing Dataset

The utilization of existing gold-standard datasets Stream-of-consciousness Essays dataset and myPersonality is an efficient approach for the personality recognition task in Indian languages. The major issue with the machine learning and deep learning approach is noisy data and incorrect labeling. Data quality affects machine learning and deep learning model performance. The cleaning and re-labeling of existing data may improve the accuracy of the model [17]. Data quality is the key issue in data management, dirty data leads to inaccurate classification and data analytics results. Data quality indicates the degree to which the data is complete, reliable, free from duplication, accurate, and timely for a given purpose. Data cleaning is the process that referred to the tasks which detect and repair errors in the data. The study [13] revealed the impact of data cleaning on machine learning models.

As discussed, translated existing dataset may contain one or more types of errors: Inconsistencies, Incorrect translation, Duplicates, Missing Values, Outliers, and Mislabels Data. In the dataset mislabels errors arise when a personality trait is wrong labeled. For mislabels errors, it is difficult to clean it without knowing domain knowledge. The study in [13] followed the flipping process of uniform class noise and pairwise class noise to repair mislabel errors. The missing values error occurs in the dataset when an empty value is stored for some trait in the dataset. The deletion and imputation methods can be used to repair missing values. The outliers are the abnormal observation and the methods Standard Deviation, Isolation Forest Method, Method Interquartile Range Method, Deletion, and Imputation have been utilized in the study [13] to detect the outliers.

4. Conclusion

Machine learning and deep learning have become more popular and important to improve personality classification results. In the direction of future research, deep learning models can provide helpful insights into the psycholinguistic feature with the availability of large and quality datasets. In this work, an existing dataset has been presented for the personality recognition task and challenges in data collection and generation in Indian languages. Furthermore, data quality issues and their solutions have been discussed for improving the existing datasets. Data quality affects the performance of machine learning and deep learning models. The generation of a new dataset for the personality recognition task in Indian languages is difficult, and so an alternative approach is required to improve the existing datasets. Cleaning the existing dataset may be the faster way to increase the accuracy of the models. In future research directions, the research may be expanded to emphasize the selection of the data cleaning methods for improving dataset quality.

References

[1] Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged SVM over BERT word embedding ensembles" in Proceedings of the ACL 2020 workshop on Widening NLP. Association for Computational Linguistics, 2022.

[2] Dargan, S., M. Kumar, M. R. Ayyagari, and G. Kumar (2019), A Survey of Deep Learning and Its Applications: A New Paradigm to Learning, Archives of Computational Methods in Engineering. doi:10.1007/s11831-019-09344-w,2019.

[3] E. Gortner, S. S. Rude, J. W. Pennebaker, "Benefits of Expressive Writing in Lowering Rumination and Depressive Symptoms", Elsevier, Behavior Therapy 37, 292–303, 2006.

[4] F. Mairesse et al., Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text, Artificial Intelligence Research, vol.30, pp 457–500,2007.

[5] Firoj Alam, Evegeny A. Stepanov, Giuseppe Riccardi , "Personality Traits Recognition on Social Network- Facebook" Computational Personality Recognition (Shared Task),2013.

[6] Golbeck, J. and Robles, C., and Turner, K., "Predicting Personality with Social Media", In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, pp. 253–262,2011.

[7] Golnoosh Farnadi, GeethaSitaramanShanuSushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos Marie-Francine Moens Martine De Cock, "Computational personality recognition in social media" User Model User-Adap Inter DOI 10.1007/s11257-016-9171.

[8] J. Golbeck, C. Robles, M. Edmondson, and K. Turner , “Predicting personality from twitter,” in Proceedings of IEEE International Conference on Social Computing, pp. 149–156,2011.

[9] Jianguo Yu, Konstantin Markov, Deep Learning based Personality Recognition from Facebook Status Updates, IEEE 8th International Conference on Awareness Science and Technology (iCAST 2017).

[10] Mehta, Y., Majumder, N., Gelbukh, A., Cambria, E., Recent trends in deep learning-based personality detection, Artificial Intelligence Review. https://doi.org/10.1007/s10462-019-09770-z Miller, G. . The smartphone psychology manifesto. Perspectives on Psychological Science, 7(3), pp 221–237. https://doi.org/ 10.1177/1745691612441215,2020.

[11] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning base document modeling for personality detection from text, IEEE Intelligent Systems, vol. 32, no. 2, pp 7479, 2017.

[12] Pennebaker, J.W. and King L.A., “Linguistic style: Language use as an individual difference. Journal of Personality and Social Psychology, 77, 1296-1312,1999.

[13] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang. CleanML: A benchmark for joint data cleaning and machine learning. CoRR, abs/1904.09483, 2019.

[14] S. Poria, A. Gelbukh, B. Agarwal, E. Cambria, N. Howard, Commonsense knowledge based personality recognition from text, in Advances in Soft Computing and Its Applications, Springer, pp 484–496.

[15] S. Han, H. Huang, and Y. Tang, Knowledge of words: An interpretable approach for personality recognition from social media, Knowledge Based Systems, vol. 194, pp 105550,2020.

[16] Tommy Tandera, Hendro, DerwinSuhartono, RiniWongso, Yen LinaPrasetio, Personality Prediction System from Facebook Users, 2nd International Conference on Computer Science and Computational Intelligence, Bali, Indonesia, pp 604-611,2017.

[17] Y. Roh, G. Heo, and S. E. Whang. A survey on data collection for machine learning: a big data - AI integration perspective. IEEE TKDE,2019.

[18] Zhancheng Ren, Qiang Shen, Xiaolei Diao, Hao Xu, A sentiment-aware deep learning approach for personality detection from text, Information Processing and Management, Article-102532,2021.

[19] https://translate.google.com/

[20] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[21] Mehl, M. R., Gosling, S. D., &Pennebaker, J. W., Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. Journal of Personality and Social Psychology, 90, 862–877,2006.